(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
25 April 2002 (25.04.2002)

PCT

(10) International Publication Number
WO 02/33626 A1

(51) International Patent Classification[7]: G06F 17/60

(21) International Application Number: PCT/US01/32178

(22) International Filing Date: 15 October 2001 (15.10.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/690,033 16 October 2000 (16.10.2000) US

(71) Applicant: **ENGAGE TECHNOLOGIES** [US/US]; 2nd floor, 100 Brickstone Square, Andover, MA 01810 (US).

(72) Inventors: **WANG, Changfeng**; 36 Fairbanks Road, Lexington, MA 02421 (US). **JAYE, Daniel, B.**; 650D Brookside Drive, Andover, MA 01810 (US).

(74) Agents: **KELLY, Edward, J.** et al.; Ropes & Gray, One International Place, Boston, MA 02110-2624 (US).

(81) Designated States *(national)*: AU, CA, JP, KR.

(84) Designated States *(regional)*: European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR).

Published:
— *with international search report*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 02/33626 A1

(54) Title: DEMOGRAPHIC PROFILING ENGINE

(57) Abstract: An apparatus and method to create demographic user profiles of communication network users. In an embodiment, the communication network is the internet, and demographic profiles are produced for internet users using URL data. A probabilistic model is created for each demographic attribute, with example demographic categories including marital status, gender, age, income, profession, industry, etc. Each probabilistic model is computationally independent of the others, and is derived by generating interest scores for a filtered URL set, identifying a discriminating URL set for each demographic attribute, and creating a probability distribution function from the respective interest scores. The probabilistic models accept an interest score and produce a corresponding demographic probability measure that indicates the probability that the respective models exist for each demographic attribute, user profiles may be maintained and updated by obtaining a specified user's URL data, filtering the data, attaching interest scores to the filtered data, inputting the interest scores to respective demographic models to produce probabilities, and updating the user profile with the newly computed probabilities.

DEMOGRAPHIC PROFILING ENGINE

Field of the Invention

The present invention relates generally to communication systems, and more particularly to user profiling in interactive communication systems such as the internet.

Description of the Prior Art

The advertising landscape has recently altered tremendously with the advent and subsequent consumer embracing of the internet. The internet, a global communication network of computers (for structure and operation, see "The Internet Complete Reference," by Harley Hahn and Rick Stout, published by McGraw-Hill, 1994, and incorporated herein by reference), has spawned such terms as "e-commerce" to describe the tremendous amount of transactions performed entirely through internet communications. Virtually every industry, and hence every competitor within each industry, is required to maintain an internet presence providing electronic goods and services, to remain competitive. The result is a large number of services and products that are available through the internet. These products and

(otherwise known as Uniform Resource Locations, or URLs) that include electronic information that may be retrieved from a designated network, and provided on a display that is connected to an internet accessible device. The majority of internet users access the internet using a personal computer, SUN workstation, handheld or laptop computer, etc., however other, non-computer devices such as televisions may provide internet access with the use of a modem.

Because of the ever-increasing numbers of internet products and services, and internet users, advertising is critical to inform users of potential websites of interest. Traditional methods of communicating website information included likewise traditional advertising techniques such as print (newspaper, magazine, etc.), radio, and television. As the internet technology has evolved, however, and the numbers of websites increase rapidly, more focused advertising techniques are required.

Internet Service Providers (ISPs) allow internet users to access the internet through the ISP computer network by allowing the user to first connect to the ISP network, whereby the ISP network then provides access to the larger series of networks known as the internet. Most ISPs receive a fee for this connection service, and the fee may be fixed, or variable depending upon use.

Alternately, Free Service Providers (FSPs) provide free Internet connection services in exchange for users viewing advertising through a software application that executes on the user's personal computer whenever the user is connected to the FSP. Although this type of advertising was once viewed as an inconvenience to the user, as website numbers increase simultaneously with FSP popularity, the benefits of such advertising opportunities are apparent to consumers (users) and advertisers.

It is predicted that internet advertising will grow to nearly $40 billion within the next eighteen to twenty-four months. With the great increase in expenditure, and the increased amount of competition, it is imperative to accurately target specific consumers most likely to take interest in the advertisement. With the rapidly increasing numbers of internet consumers, it is extremely difficult to selectively identify interested users.

Prior art discloses an electronic advertising system wherein advertisements are placed on a system using the internet or telephone by entering data comprising a personal profile, and advertisers reply based upon the entered profile. Other prior art presents a system and method for providing customized advertisements across an interactive communication

system wherein consumer profiles are integrated with content providers to generate advertisement requests. The consumer reaction and response to the targeted advertisements are tracked. Unfortunately, the prior art systems require consumers to input and continually update their profiles to maintain accurate profile information. It is increasingly difficult to obtain information from consumers voluntarily as consumers seek to protect their privacy and limit an ever-increasing numbers of unsolicited mailings, emails, and telephone calls; however, demographic data is essential to properly identifying advertising targets.

There is currently not a sufficiently accurate system or method to obtain demographic data about communication system users without continually requiring specific data from the communication systems users.

What is needed is a system and method that accurately predicts communication system user demographics.

Summary of the Invention

The methods and systems disclosed herein provide an apparatus and method to create demographic profiles of users of an interactive communications network. In one embodiment, the communications system is the internet, and the invention derives probabilistic models of

demographic attributes using communications systems data from users with known demographic profiles. In one embodiment, communications systems data may include URL data, and in an embodiment, the URL data may be provided by an Internet Service Provider (ISP). Additionally, demographic attributes may include marital status, gender, age, income, profession, industry, etc.

In one embodiment of the methods and systems disclosed herein, each probabilistic demographic model may computationally independent of the other models, and each model may be derived by identifying a discriminating URL set for each demographic attribute, generating interest scores for each discriminating URL, and creating a probability distribution function from the respective interest scores.

In one embodiment, the methods and systems disclosed herein determine a set of discriminating URLs for each demographic attribute, wherein the discriminating URLs maintain a specified likelihood of relating to a particular demographic attribute. In an embodiment, the methods and systems compute interest scores that relate the interest level of a particular user to a particular URL. In one embodiment, the methods and systems create a probabilistic distribution function for each demographic attribute from the

interest scores of the discriminating URLs for that demographic attribute.

The systems and methods described herein may create a set of key word indices from the content of a URL and query information in the URL. In one embodiment, the methods and systems may utilize the key word indices to perform content association and derive a set of surrogate discriminating URLs. In an embodiment, the key word indices and surrogate URLs may be used as an alternative to discriminating URLs in situations where discriminating URLs are not available.

In one embodiment, the methods and systems described herein may generate a predictive model for each demographic attribute, where the input to the predictive model may be an interest score for a specified URL, and the output may be a probability indicating whether a specified user adheres to the particular demographic attribute.

In an embodiment, interest scores input to a predictive model may be computed from websites that are deemed to be discriminating URLs for a particular attribute, and the interest scores may be generated using information including browsing duration, timestamp, content or query generated key word indices, and frequency of visit to the particular website.

The methods and systems disclosed herein may maintain and update user profiles by employing probabilistic models that exist for relevant demographic attributes, wherein the user profiles may be maintained and updated by obtaining a specified user's URL data, attaching interest scores to the discriminating URL data, inputting the interest scores to respective demographic models to produce demographic probabilities, and updating the user profile with the newly computed probabilities.

It is an aspect of the invention to continually monitor and update the demographic models, and rebuild the models when the models no longer satisfy a predetermined criteria.

Other objects and advantages of the present invention will become more obvious hereinafter in the specification and drawings.

Brief Description of the Drawings

A more complete understanding of the invention and many of the attendant advantages thereto will be appreciated as the same becomes better understood by reference to the following detailed description when considered in conjunction with the accompanying drawings, wherein like reference numerals refer to like parts and wherein:

FIG. 1 is an architectural and functional block diagram of a system practicing the principles for generating the demographic profile engine invention, including a model builder and a profile component;

FIG. 2 is a functional block diagram of the FIG. 1 model builder; and,

FIG. 3 is a functional block diagram representing demographic profile operation and updating.

Best Mode For Carrying Out The Invention

To provide an overall understanding of the invention, certain illustrative embodiments will now be described; however, it will be understood by one of ordinary skill in the art that the systems described herein can be adapted and modified to provide systems for other suitable applications and that other additions and modifications can be made to the invention without departing from the scope hereof.

Referring now to FIG. 1, there is shown an architectural block diagram for illustrating the components of the invention as practiced herein and known as a demographic profiling engine **10**. The demographic profiling engine **10** as presented shall be applied to the interactive communications network known commonly as the internet, however, the invention herein is not so limited, and may be applied to any

communications network. The demographic profile engine **10** creates probabilistic demographic models, applies user data to the models, and generates and updates user profiles from the model outputs. For discussion purposes, the illustrated demographic profiling engine **10** may therefore be viewed as having two major components: a Model Builder **12** and a Profiling Component **14**.

Referring to FIG. 1, the illustrated Model Builder **12** creates a probabilistic model for each demographic attribute, thereby making each demographic model computationally independent of any other demographic model. In the FIG. 1 system, the individual demographic models **16** are provided to the Profiling Component **14**, wherein user data **18** is selectively applied to the distinct demographic models **16** to produce demographic statistics for a respective internet user. The demographic statistics are combined with existing user demographic profile information to update **20** the respective user profile for storage in the user profile database **22**.

FIG. 2 details the components and processing of the illustrated Model Builder **16** from FIG. 1. The illustrated Model Builder **16** accesses a database having demographic data for known users, wherein in this illustration, the database shall herein be referenced as

a Data Set Server, or DSS **30**. For the internet application presented herein, the DSS **30** contains user-specific information for known, different internet users, wherein the information may include historical clickstream data of Uniform Resource Locations (URLs), cookie IDs, browsing duration, timestamps (i.e., recency), content or query keyword indices, frequency of visit, and demographic profile data. The content or query indices may be generated using well known techniques that utilize URL content and server log query data.

In the illustrated embodiment, the DSS data is communications systems data that is derived primarily from a survey of known communications system users; and, in the internet application discussed herein, the data may be derived from such sources as an Internet Service Provider (ISP) or an offline data source such as a marketing data source, and those with ordinary skill in the art will recognize that the invention herein is not limited to the source of data that, in the illustrated embodiment, is contained within the DSS **30**. Additionally, DSS **30** demographic profile data that accompanies the communications systems data, may include marital status, gender, age, income, profession, and industry, however the invention herein is also not limited to this specified list of demographic attributes

or other known user information, and may be expanded or reduced accordingly without affecting the invention.

The illustrated DSS **30** provides user-specific information to a Mining Database **32** that is responsible, according to the FIG. 2 embodiment, for filtering the DSS **30** data, including the URLs and keyword indices. In an embodiment, URL filtering reduces the number of URLS, and such reduction may be because a URL is redundant, or deemed irrelevant or insignificant. Criteria for URL determining irrelevant or insignificant data may include low frequency of visit, URLs not visited recently, and URLs that do not meet some predetermined criteria, although the URL filtering process may be adapted to the application and does not limit the present invention.

In one embodiment, after the URL filtering is performed, the remaining filtered URLs are presented to one of many well-known algorithms to extract keywords from the filtered URLs and generate a keyword index for each filtered URL. The keyword indices for the filtered URLs may then be compared against URLs that are otherwise not associated with a particular user, to identify URLs that are closely associated with the filtered URLs. Such closely associated URLs are identified herein as surrogate URLs. As will be disclosed herein, surrogate URLs may be utilized to

increase the amount of predictive data available to the models.

Identification of surrogate URLs may be performed in many different ways, and the invention herein is not limited by the technique or method for comparing and/or associating URLs. In one sample embodiment, frequency of terms and other statistic measures including inverse document frequency (IDF) may be utilized to associate a keyword index of one URL with a keyword index of a different URL.

In the illustrated system, a parallel process as described for filtering URLs is performed for query data. Clickstream data relating to queries of a particular user, is identified and analyzed to generate a keyword index for each query. Query indices may then be filtered by content to reduce redundant or otherwise irrelevant data.

After the illustrated Mining Database **32** generates the URLs and query keyword indices, it computes an interest score for the filtered URLs and/or query keyword indices. Interest scores may represent an internet user's interest in a particular URL or query, and in one embodiment, an interest score is computed as a number between zero and one, with one representing maximum interest of a particular user to a URL or query. In an embodiment, URL interest scores may be a function

of URL page length, URL visit duration, URL frequency of visit, and/or the time since the URL was last visited (recency). Those with ordinary skill in the art will recognize that interest scores may be computed using a single or multiple parameters, using various methodologies, and the invention herein is not limited by the interest score computation method. Although in the illustrated embodiment, interest scores are also computed for query indices, it should be noted that query indices may not be available for all users. In the illustrated system, interest scores are computed for discriminating URLs and query keyword indices only to the extent that such data is available.

Once the illustrated Mining Database **32** attaches an interest score to each filtered URL and/or query keyword index, the FIG. 2 Mining Database **32** provides the filtered URLs and/or query keyword indices and associated interest scores to a Feature Selector **34**. The illustrated Feature Selector **34** is responsible for identifying those Mining Database filtered URLs and/or query keyword indices that should be applied to each of the different demographic categories. In the illustrated embodiment, this process is known as defining a discriminating feature set, and an individual discriminating feature set is derived for each demographic attribute. In the illustrated system, the

discriminating feature set may include discriminating URLs, discriminating query keyword indices, and discriminating category subject matter (i.e., "sports", "education", etc.). The illustrated system uses well-known techniques to generate an interest score in a category subject matter for a (known or unknown) user when appropriate. As mentioned previously, every (known or unknown) user may not provide all of a discriminating URL, discriminating query keyword index, and discriminating category subject matter.

In the illustrated embodiment of FIG. 2, the Feature Selector **34** identifies discriminating URLs probabilistically. For example, in one embodiment, for a given URL denoted as "l", the Mining Database **32** may provide a corresponding interest score X_l; however, there is additionally a set of demographic attributes A, which may be, for example, a set of demographic attributes including age, marital status, gender, income, profession, industry, etc., although the invention herein is not so limited to the demographic set.

The illustrated Feature Selector **34** computes, for each of the known users, and for each element within the demographic set, a probability density function (PDF) expressed as the conditional probability mathematically expressed by equation (1).

$$P(A = a \mid X_{l_1}(t), X_{l_2}(t), X_{l_3}(t),..., X_{l_m}(t), X_{l_1}(t-1), X_{l_2}(t-1),..., X_{l_n}(t-1),...)$$

, (1)

where a specific user visited a subset of m URLs, designated as l_i, i=1 to m, during a communications or internet session t, that generated corresponding interest scores of $X_{l_1}(t)$; and, the known user visited a set of n URLs during internet session t-1 with corresponding interest scores, and the "..." at the end of equation (1) denotes previous visiting history (i.e., times t-2, t-3, etc.) URL interest scores for this user. The probability given by equation (1) shall be referred to herein as the demographic score, or "dscore", of attribute value A=a. The dscore represents the probability that a particular user is a member of the demographic model.

In the embodiment of the invention presented herein, every URL visited by a particular (known or unknown) user may not contribute to the dscore of a particular attribute, A. In the illustrated embodiment of FIG. 2, a URL, m, may be a discriminating URL for a given attribute A if the interest score for that URL is a "relevant" feature in computing the dscore. To determine whether a feature is a relevant feature with respect to a given attribute A, the illustrated system of FIG. 2 computes the Kullback-Leibler distance between the PDF computed using the interest score $X_{l_m}(t)$ (equation

2) and the PDF computed without using the interest score $X_{Im}(t)$ (equation 3), for internet sessions t, t-1, etc., of a known user:

$$P(A=a \mid X_{I1}(t), X_{I2}(t), X_{I3}(t), \ldots, X_{Im}(t), X_{I1}(t-1), X_{I2}(t-1), \ldots, X_{In}(t-1), \ldots) \quad (2)$$

$$P(A=a \mid X_{I1}(t), X_{I2}(t), X_{I3}(t), \ldots, X_{Im-1}(t), X_{I1}(t-1), X_{I2}(t-1), \ldots, X_{In}(t-1), \ldots) \quad (3)$$

If the Kullback-Leibler difference between the two PDFs as computed for internet sessions t, t-1, etc., by equations (2) and (3), is zero, then the URL with corresponding interest score $X_{Im}(t)$, is deemed to be a non-discriminating URL for attribute A. Alternately, if the Kullback-Leibler difference is not zero, the URL is a discriminating URL for attribute A.

The determination of discriminating query keyword indices and category subject matter information is also developed probabilistically in the illustrated system.

In the FIG. 2 system, the discriminating URLs, associated surrogate URLs, and query keyword indices for each demographic attribute are maintained in a "hash table" that identifies discriminating information for the demographic models. In the illustrated systems, the hash table, derived from discriminating URLs, etc. from known users, is utilized to identify discriminating data for unknown users. By including surrogate information in the hash tables, the amount of information that may

be utilized to profile unknown users, increases. For example, although one website for selling compact disks (CDs) may be a discriminating URL for a particular demographic model, surrogate URLs to this discriminating URL may also be identified, wherein the surrogate URLs are other URLs that also sell CDs. The illustrative methods and systems disclosed herein may therefore generate essentially equivalent profiling information from either the discriminating URL, or one if its surrogate URLs.

The FIG. 2 Feature Selector **34** transfers the discriminating feature set (i.e., URLs, query keyword indices, and category subject matter) and respective interest scores for each demographic attribute, to a Model Constructor and Validator **36**. The illustrated Model Constructor and Validator **36** constructs a probabilistic model of each demographic attribute from the demographic attribute's respective feature set and associated interest scores. In the illustrated embodiment, the model construction is performed by invoking a training algorithm that is related to a selected learning algorithm. Those with ordinary skill in the art will recognize that learning algorithm selection may be based upon the generated model, and the training algorithm utilizes information from discriminating URLs. Although in one embodiment, model

training may be performed in a batch process that is computed as a background task on a daily or weekly basis, the models in the FIG. 2 system that are derived from the training are deployed on the internet for real-time applications such as advertisement placement or dynamic presentation of web pages, and it is therefore important that the models have a computation time that is compatible with such real-time applications.

Although computational efficiency may be a consideration when selecting a learning algorithm from which a model is derived, other considerations may include an unavailability during the current session, of discriminating URLs or surrogate URLs. Similarly, a model may receive varying input dimensions and may not always receive every input value. A learning algorithm should therefore be selected to optimize the potentially varying situations of each application.

The illustrated Model Constructor and Validator **36** is also responsible for validating the models. In the illustrated embodiment, the models are validated immediately after training, and on a continual basis thereafter. In the illustrated embodiment, the models are validated using a known set of user data, and validation is performed by ensuring that the error between the known user profiles and the model outputs is within a predetermined error rate. If the error rate

for any given model is not within the predetermined error rate, that model may be rebuilt using the process described by FIG. 2. When models are rebuilt, the demographic data in the DSS **30** may be altered or augmented when compared to the previous time that the particular model was built. As mentioned previously, demographic models in the illustrated embodiment are independent, and may be continually rebuilt independently of the other models.

Although the illustrated Model Constructor and Validator **36** continually evaluates each demographic model against a known user data group to ensure that each demographic model may be within the predetermined error rate, in other embodiments, the model evaluation and rebuilding may be performed at fixed intervals, and models may be rebuilt with respect to the rebuilding of other models.

As an example, consider a system practicing the invention, wherein the system is based upon the Naïve Bayes Network. Although this example is provided for illustrative purposes, those with ordinary skill in the art will recognize that other predictive algorithms, including but not limited to Neural Networks, Decision Tree algorithms, etc., may be substituted without departing from the scope of the invention. Using the notation previously provided, a recursive equation for

computing dscores using Bayes Rule may be expressed as equation (4):

$$P(A=a \mid X_{l1}(t), X_{l2}(t), X_{l3}(t), \ldots, X_{lm}(t), X_{l1}(t-1), X_{l2}(t-1), \ldots, X_{ln}(t-1), \ldots) = \frac{1}{c(t)} \prod_{i=1}^{m} \frac{P(X_i(t) \mid a)}{P(X_i(t-1) \mid a)} P(A=a \mid X_{l1}(t-1), X_{l2}(t-1), \ldots, X_{ln}(t-1), \ldots)$$

(4)

where c(t) is a normalization factor. If the logarithm of equation (4) is taken, the computation for dscores (i.e., $P(A=a \mid X_{l1}(t), X_{l2}(t), X_{l3}(t), \ldots, X_{lm}(t), X_{l1}(t-1), X_{l2}(t-1), \ldots, X_{ln}(t-1), \ldots)$) may be expressed as a recursive update equation from one internet session to the next internet session, indicated by equation (5):

$$d_a(t) = d_a(t-1) + \sum_{i=1}^{m} \{p_{ai}(t) - p_{ai}(t-1)\},$$

(5)

where $p_{ai}(t) \cong \log(P\{Xi(t) \mid A=a\}$ and $d_a(0)=\log(P\{A=a\})$, and the function $p_{ai}(t)$ is provided by the training algorithm of the Bayes Network in a batch process, while the a priori probability is derived from the training process and stored as a constant. Some of the advantages of this Bayes Network illustrative system may include the ability to easily handle varying input dimensions, very low computational complexity in training, low run-time complexity, and the ability to implement parallel implementation. One disadvantage may be that the algorithm may make unrealistic assumptions on the data

set distribution. This problem, however, may be approached using a well-known "boost" algorithm that enhances the existing classifier.

Continuing the Bayes example, computating the Kullback-Liebler distance reduces to computing the mutual information, $I(X_1, A)$ between a URL, X_1, and an attribute, A. It may be determined in such a system that a URL, X_1, is discriminating if:

$$I(X_1, A) = H(X_1) - H(X_1|A) > 0, \quad (6),$$

and otherwise it is non-discriminating, where:

$$H(X_i) = \sum_x P(X_i = x) * \log(P(X = x)) \quad (7)$$

and,

$$H(X_i|A) = \sum_{x,a} \Pi_a * P\{X_i = x | A = a\} * \log(P(X = x | A = a)) \quad (8)$$

where, Π_a is the empirical version of P(A=a).

An algorithm for determining discriminating features may thus include the steps of training the Bayes classifier on attributes of a training data set; setting the set of discriminating URLs to zero for a particular attribute; for each URL, computing the mutual information $I(X_1, A)$; and, including the URL in the discriminating set of URLs for the attribute if $I(X_1, A)$ is greater than zero. Once a discriminating URL set is obtained for each attribute, the models may be built and validated.

Referring now to FIG. 3, there is shown a functional block diagram of the illustrated Profiling Component **14** of FIG. 1. As mentioned previously, in the illustrated embodiment, the Profiling Component **14** utilizes the validated demographic models provided by the Model Builder **12**, to generate and update demographic profiles for users that shall be known herein as "unknown users", to be distinguished from "known users" from whom data was obtained to determine the models. The FIG. 3 Profiling Component **14** accepts as input a data set **40** that in the illustrated embodiment, may include a cookie ID, a URL clickstream, a browsing duration, timestamp (recency), query keyword index, frequency of visit, etc., for an internet session t and a particular unknown user X, however the invention is not limited to such information, and other data set information may be included or eliminated without departing from the scope of the invention.

In the illustrated system, the data from the unknown user is analyzed and filtered to determining discriminating data. In the system of FIG. 3, the URL data is filtered to determine those URLs that are the same as a discriminating URL for a demographic model, or the same as a surrogate URL associated with a demographic model. Similarly, a keyword index may be generated for any query information received from the

unknown user, and the keyword index may be compared to discriminating keyword indices. URLs or query information that is not deemed to match or otherwise associate with previously defined discriminating (i.e., including surrogate) information may be eliminated.

In the FIG. 3 system, an interest score is computed for each remaining URL **42** and query keyword index in the data set **40**, and this interest score computation **42** may be computed using the same process used in the Model Builder **12**. The interest scores for each URL of the illustrated system are then applied to each demographic attribute model **44** to output a probability that user X is a member of that demographic group, the probability being otherwise known as a "dscore" for each attribute. Any previously existing user profile for user X may then be retrieved from a historical user profile database **46** or other location and updated with the latest dscores. The updating may be performed using any well-known data filtering technique, learning algorithm, etc. Those with ordinary skill in the art will recognize that if a user profile does not exist, a new user profile may be created. The updated or new user profile may similarly be saved for later retrieval by an advertising system, another update, etc. The profile database **46** may be arranged in any manner to facilitate the application utilizing the user profiles stored therein, and those

with ordinary skill in the art will recognize that the database **46** may be replaced with any other storage mechanism that performs the functions described herein for storing and allowing retrieval of user profile information. For example, the database may be may be realized using any database or relational database system, such as Oracle 8, SQL, MySQL, or any other database system.

An advantage of the present invention over the prior art is that independent demographic models may be computed and applied to communications users to generate a user profile without requiring data input by the users.

What has thus been described is an apparatus and method to create demographic user profiles of communication network users. In one embodiment, the communication network is the internet, and demographic profiles are generated for internet users using URL data. A probabilistic model is created for each demographic attribute, with example demographic categories including marital status, gender, age, income, profession, industry, etc. In the illustrated embodiment, each probabilistic model is computationally independent of the others, and is derived by generating interest scores for a filtered URL set, identifying a discriminating URL set for each demographic attribute,

and creating a probability distribution function from the respective interest scores. The probabilistic models accept an interest score and produce a corresponding demographic probability measure that indicates the probability that the respective user is within the demographic group. Once probabilistic models exist for each demographic attribute, user profiles may be maintained and updated by obtaining a specified user's URL data, filtering the data, attaching interest scores to the filtered data, inputting the interest scores to respective demographic models to produce probabilities, and updating the user profile with the newly computed probabilities.

Although the present invention has been described relative to a specific embodiment thereof, it is not so limited. Obviously many modifications and variations of the present invention may become apparent in light of the above teachings. For example, the block diagrams and functions related thereto are merely representative and functionality may be combined without departing from the scope of the invention. Although in the illustrated embodiment, the demographic attribute models were based upon the same prediction algorithm, different demographic attribute models may be based upon different prediction algorithms. The attribute models may be updated on fixed intervals or, as in the illustrated

embodiment, intervals that are independently computed with respect to the other models. The invention may be applied to any communications network. The functionality of the different components of the illustrated embodiments may be otherwise partitioned or combined without affecting the scope of the invention.

Many additional changes in the details, materials, steps and arrangement of parts, herein described and illustrated to explain the nature of the invention, may be made by those skilled in the art within the principle and scope of the invention. Accordingly, it will be understood that the invention is not to be limited to the embodiments disclosed herein, may be practiced otherwise than specifically described, and is to be understood from the following claims, that are to be interpreted as broadly as allowed under the law.

We claim:

1. A method for producing a demographic attribute model, comprising,

obtaining known demographic data associated to communications system data,

computing interest scores relating an interest level to the communications system data,

correlating the communications system data to a demographic attribute, and,

generating a probabilistic model for the demographic attribute using the interest score from the correlated communications systems data.

2. A method according to claim 1, wherein obtaining a set of known demographic data associated to communications systems data further comprises receiving demographic data with corresponding URL clickstream data from an Internet Service Provider (ISP).

3. A method according to claim 1, further comprising filtering the communications system data.

4. A method according to claim 3, wherein filtering further comprises reducing redundancy.

5. A method according to claim 3, wherein filtering further comprises reducing statistically unreliable data.

6. A method according to claim 1, further comprising generating a keyword index for the filtered communications system data.

7. A method according to claim 6, further comprising associating the keyword index with communications system data that lacks an association with known demographic data.

8. A method according to claim 7, wherein associating a keyword index data further comprises obtaining a set of surrogate URLs.

9. A method according to claim 7, wherein associating a keyword index data further comprises utilizing at least one of term frequency, inverse document frequency, or content association.

10. A method according to claim 1, wherein correlating the communications system data to a demographic attribute further comprises,

designating a selected communications data element,

computing a first conditional probability of the demographic attribute given the interest scores for the communications system data including the selected communications data element,

computing a distinct second conditional probability of the demographic attribute given the interest scores of the communications system data without including the selected communications system data;

differencing the first and the distinct second conditional probabilities;

correlating the selected communications system data to the demographic attribute when the difference is zero.

11. A method according to claim 1, wherein generating a probabilistic model further comprises executing a training algorithm.

12. A method according to claim 11, wherein executing a training algorithm further comprises selecting a probabilistic model.

13. A method according to claim 1, further comprising validating the model.

14. A method according to claim 13, wherein validating the model further comprises comparing the model against known demographic data.

15. A method according to claim 1, further comprising monitoring the model for accuracy.

16. A method according to claim 15, wherein monitoring the model further comprises statistically evaluating the model at defined intervals.

17. A method according to claim 15, further comprising rebuilding the model upon finding that the model is not within a predetermined error rate.

18. A method according to claim 17, wherein rebuilding the model further comprises,

augmenting the known demographic data, and,

iteratively returning to obtaining known demographic data associated to communications system data.

19. A method for producing a demographic profile for a user of a communications network, comprising

collecting data associated to the user from the communications network,

applying the collected data to at least one probabilistic model that corresponds to a demographic attribute, and,

generating the demographic profile from outputs from the probabilistic models.

20. A method according to claim 19, wherein collecting data further comprises extracting at least one of a cookie ID, a URL clickstream, a browsing duration, a content keyword index, a query keyword index, a timestamp, or a frequency of visit data.

21. A method according to claim 19, wherein collecting data further comprises obtaining internet data.

22. A method according to claim 19, wherein applying the communications data further comprises filtering the communications data.

23. A method according to claim 22, wherein filtering the communications data further comprises reducing redundant data.

24. A method according to claim 22, wherein filtering the communications data further comprises reducing URLs as a function of a browse duration not exceeding a specified interval.

25. A method according to claim 22, wherein filtering the communications data further comprises reducing URLs lacking an association with at least one demographic model.

26. A method according to claim 19, further comprising computing an interest score that relates the user's interest level to the collected data.

27. A method according to claim 26, wherein computing an interest score further comprises associating at least one of a page length, a user duration, recency, or a frequency of visit, to a Uniform Resource Location (URL).

28. A method according to claim 26, wherein computing an interest score further comprises associating at least one of a page length, a user duration, recency, or a frequency of visit, to a keyword index.

29. A method according to claim 19, wherein applying the communications data to the at least one probabilistic model further comprises inputting information associated with the communications data to a Naïve Bayes Network.

30. A method according to claim 19, wherein applying the communications data to the at least one probabilistic model further comprises inputting information associated with the communications data to a Decision Tree Algorithm.

31. A method according to claim 19, wherein applying the communications data to the at least one probabilistic model further comprises inputting information associated with the communications data to a Support Vector Machine.

32. A method according to claim 19, wherein applying the communications data to the at least one probabilistic model further comprises creating a Neural Network for at least one demographic attribute.

33. A method according to claim 19, wherein utilizing the probabilistic model outputs further comprises,

retrieving a previously produced demographic profile for the user, and,

incorporating the probabilistic model outputs with the previously produced demographic profile.

34. A method according to claim 33, further comprising storing the demographic profile to a database.

35. A method according to claim 33, wherein incorporating the probabilistic model outputs further comprises applying a learning algorithm.

36. A method according to claim 19, wherein applying the communications data further comprises associating the communications data to discriminating data for the demographic models.

37. A method according to claim 19, further comprising generating at least one demographic model.

38. A method according to claim 37, wherein generating at least one demographic model further comprises,

obtaining known user communications systems data, and,

associating the known user communications systems data to a demographic attribute.

39. A method according to claim 38, wherein associating the known user communications systems data further comprises filtering the known user communications data.

40. A method according to claim 38, wherein associating the known user communications data further comprises generating a keyword index for the known user communications data.

41. A method according to claim 38, wherein associating the known user communications data further comprises generating surrogate communications data.

42. A method according to claim 41, wherein generating surrogate data further comprises,

generating keyword indices the known user communications data, and

associating the keyword indices of the known user communications data with keyword indices of other communications data.

43. A method according to claim 42, wherein associating the keyword indices further comprises utilizing at least one of term frequency, inverse document frequency, or content association.

44. A computer product disposed on a computer readable medium, for producing a demographic attribute model, the computer products comprising instructions to cause a processor to,

obtain known demographic data associated to communications system data,

compute interest scores relating an interest level to the communications system data,

correlate the communications system data to a demographic attribute, and,

generate a probabilistic model for the demographic attribute using the interest score from the correlated communications systems data.

45. A computer product according to claim 44, wherein instructions to obtain a set of known demographic data associated to communications systems data further comprise instructions to receive demographic data with corresponding URL clickstream data from an Internet Service Provider (ISP).

46. A computer product according to claim 44, further comprising instructions to filter the communications system data.

47. A computer product according to claim 46, wherein instructions to filter further comprise instructions to reduce redundancy.

48. A computer product according to claim 46, wherein instructions to filter further comprise instructions to reduce statistically unreliable data.

49. A computer product according to claim 44, further comprising instructions to produce a keyword index for the filtered communications system data.

50. A computer product according to claim 49, further comprising instructions to associate the keyword index with communications system data that lacks association with known demographic data.

51. A computer product according to claim 50, wherein instructions to associate a keyword index data further comprises instructions to obtain a set of surrogate URLs.

52. A computer product according to claim 50, wherein instructions to associate a keyword index data further comprise instructions to utilize at least one of term frequency, inverse document frequency, or content association.

53. A computer product according to claim 44, wherein instructions to correlate the communications system data to a demographic attribute further comprise instructions to,

designate a selected communications data element,

compute a first conditional probability of the demographic attribute given the interest scores for the communications system data including the selected communications data element,

compute a distinct second conditional probability of the demographic attribute given the interest scores of the communications system data without including the selected communications system data;

difference the first and the distinct second conditional probabilities;

correlate the selected communications system data to the demographic attribute when the difference is zero.

54. A computer product according to claim 44, wherein instructions to produce a probabilistic model further comprise instructions to execute a training algorithm.

55. A computer product according to claim 54, wherein instructions to execute a training algorithm further comprise instructions to select a probabilistic model.

56. A computer product according to claim 44, further comprising instructions to validate the model.

57. A computer product according to claim 56, wherein instructions to validate the model further comprises instructions to compare the model against known demographic data.

58. A computer product according to claim 44, further comprising instructions to monitor the model for accuracy.

59. A computer product according to claim 58, wherein instructions to monitor the model further comprise

instructions to statistically evaluate the model at defined intervals.

60. A computer product according to claim 58, further comprising instructions to rebuild the model upon finding that the model is not within a predetermined error rate.

61. A computer product according to claim 60, wherein instructions to rebuild the model further comprise instructions to,

augment the known demographic data, and,

iteratively return to obtain known demographic data associated to communications system data.

62. A computer product disposed on a computer readable medium, for producing a demographic profile for a user of a communications network, the computer products comprising instructions for causing a processor to,

collect data associated to the user from the communications network,

apply the collected data to at least one probabilistic model that corresponds to a demographic attribute, and,

produce the demographic profile from outputs from the probabilistic models.

63. A computer product according to claim 62, wherein instructions to collect data further comprise instructions to extract at least one of a cookie ID, a URL clickstream, a browsing duration, a content keyword index, a query keyword index, a timestamp, or a frequency of visit data.

64. A computer product according to claim 62, wherein instructions to collect data further comprise instructions to obtain internet data.

65. A computer product according to claim 62, wherein instructions to apply the communications data further comprise instructions to filter the communications data.

66. A computer product according to claim 65, wherein instructions to filter the communications data further comprise instructions to reduce redundant data.

67. A computer product according to claim 65, wherein instructions to filter the communications data further comprise instructions to eliminate URLs based upon a browse duration not exceeding a specified interval.

68. A computer product according to claim 65, wherein instructions to filter the communications data further comprise instructions to eliminate URLs not associated with at least one demographic model.

69. A computer product according to claim 62, further comprising instructions to compute an interest score that relates the user's interest level to the collected data.

70. A computer product according to claim 69, wherein instructions to compute an interest score further comprise instructions to associate at least one of a page length, a user duration, recency, or a frequency of visit, to a Uniform Resource Location (URL).

71. A computer product according to claim 69, wherein instructions to compute an interest score further comprise instructions to associate at least one of a page length, a user duration, recency, or a frequency of visit, to a keyword index.

72. A computer product according to claim 62, wherein instructions to apply the communications data to the at least one probabilistic model further comprise

instructions to input information associated with the communications data to a Naïve Bayes Network.

73. A computer product according to claim 62, wherein instructions to apply the communications data to the at least one probabilistic model further comprise instructions to input information associated with the communications data to a Decision Tree Algorithm.

74. A computer product according to claim 62, wherein instructions to apply the communications data to the at least one probabilistic model further comprise instructions to input information associated with the communications data to a Support Vector Machine.

75. A computer product according to claim 62, wherein instructions to apply the communications data to the at least one probabilistic model further comprise instructions to create a Neural Network for at least one demographic attribute.

76. A computer product according to claim 62, wherein instructions to utilize the probabilistic model outputs further comprise instructions to,

retrieve a previously produced demographic profile for the user, and,

incorporate the probabilistic model outputs with the previously produced demographic profile.

77. A computer product according to claim 76, further comprising instructions to store the demographic profile to a database.

78. A computer product according to claim 76, wherein instructions to incorporate the probabilistic model outputs further comprise instructions to applying a learning algorithm.

79. A computer product according to claim 62, wherein instructions to apply the communications data further comprise instructions to associate the communications data to discriminating data for the demographic models.

80. A computer product according to claim 62, further comprising instructions to produce at least one demographic model.

81. A computer product according to claim 80, wherein instructions to produce at least one demographic model further comprise instructions to,

obtain known user communications systems data, and,

associate the known user communications systems data to a demographic attribute.

82. A computer product according to claim 81, wherein instructions to associate the known user communications systems data further comprise instructions to filter the known user communications data.

83. A computer product according to claim 81, wherein instructions to associate the known user communications data further comprise instructions to produce a keyword index for the known user communications data.

84. A computer product according to claim 81, wherein instructions to associate the known user communications data further comprise instructions to produce surrogate communications data.

85. A computer product according to claim 84, wherein generating surrogate data further comprises,

generating keyword indices the known user communications data, and

associating the keyword indices of the known user communications data with keyword indices of other communications data.

86. A computer product according to claim 85, wherein associating the keyword indices further comprises utilizing at least one of term frequency, inverse document frequency, or content association.

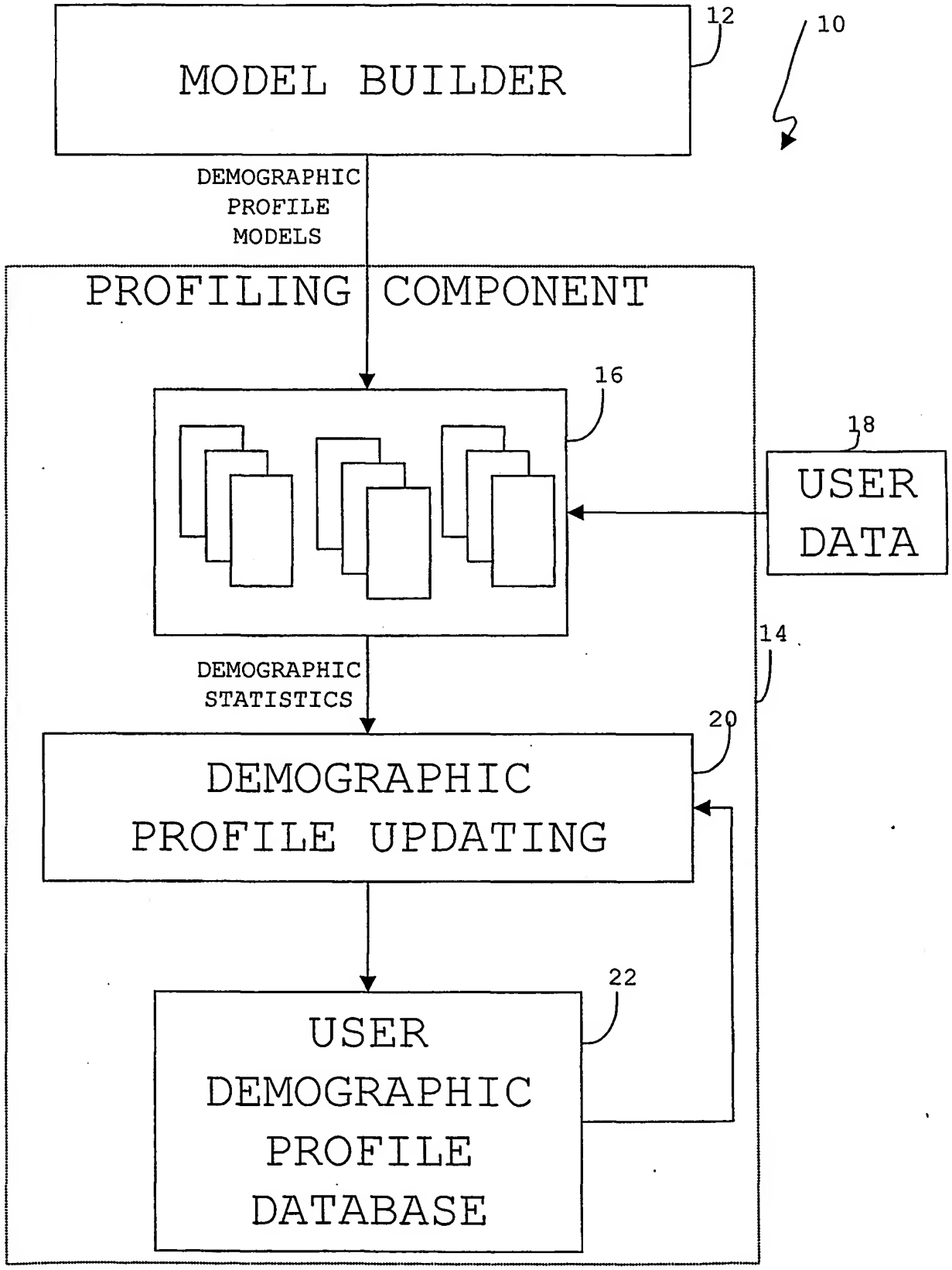
MODEL BUILDER
12
10
DEMOGRAPHIC PROFILE MODELS
PROFILING COMPONENT
16
18
USER DATA
14
DEMOGRAPHIC STATISTICS
20
DEMOGRAPHIC PROFILE UPDATING
22
USER DEMOGRAPHIC PROFILE DATABASE

FIG. 1

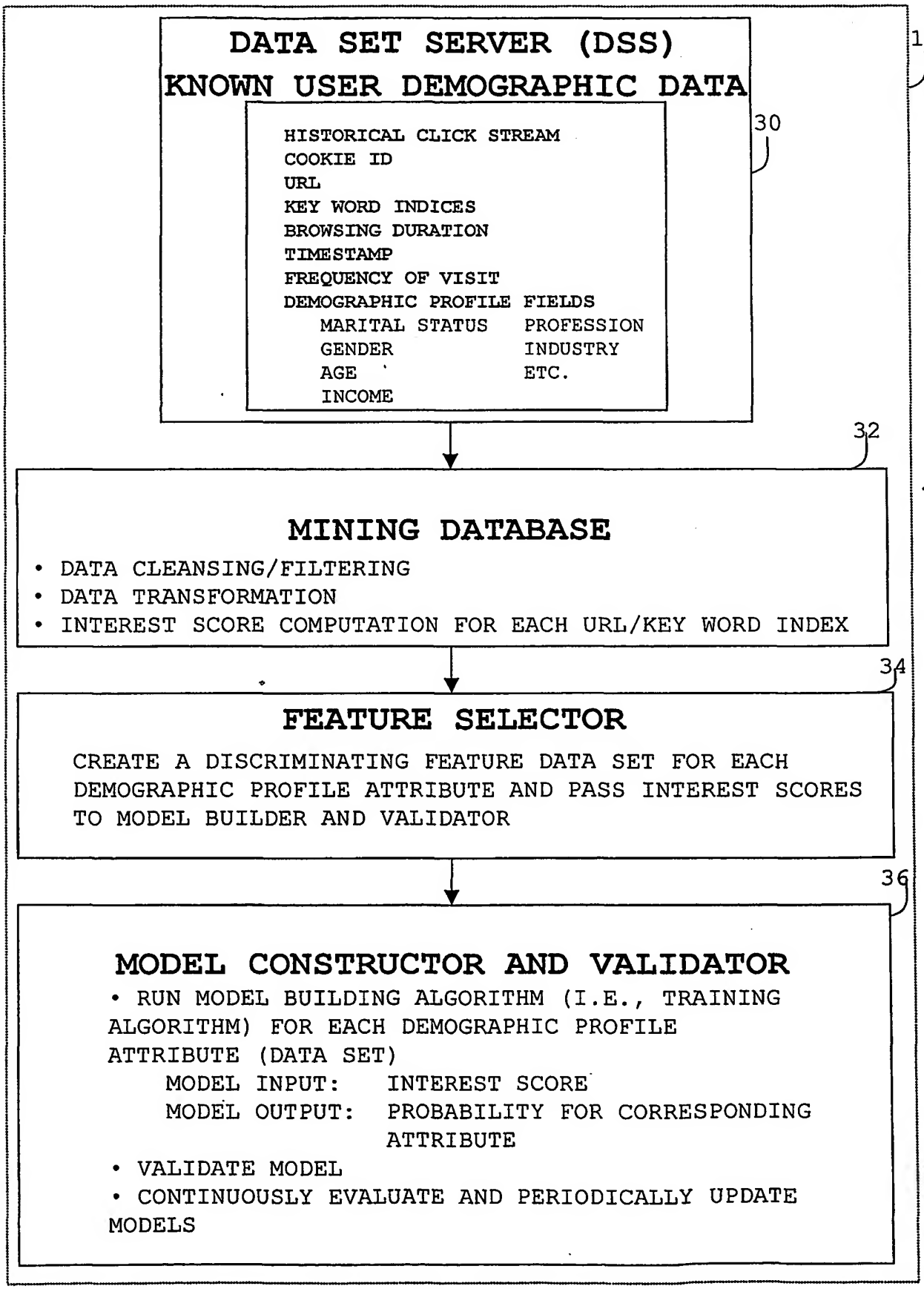
DATA SET SERVER (DSS)
KNOWN USER DEMOGRAPHIC DATA
16
30
HISTORICAL CLICK STREAM
COOKIE ID
URL
KEY WORD INDICES
BROWSING DURATION
TIMESTAMP
FREQUENCY OF VISIT
DEMOGRAPHIC PROFILE FIELDS
MARITAL STATUS
GENDER
AGE
INCOME
PROFESSION
INDUSTRY
ETC.
32
MINING DATABASE
• DATA CLEANSING/FILTERING
• DATA TRANSFORMATION
• INTEREST SCORE COMPUTATION FOR EACH URL/KEY WORD INDEX
34
FEATURE SELECTOR
CREATE A DISCRIMINATING FEATURE DATA SET FOR EACH DEMOGRAPHIC PROFILE ATTRIBUTE AND PASS INTEREST SCORES TO MODEL BUILDER AND VALIDATOR
36
MODEL CONSTRUCTOR AND VALIDATOR
• RUN MODEL BUILDING ALGORITHM (I.E., TRAINING ALGORITHM) FOR EACH DEMOGRAPHIC PROFILE ATTRIBUTE (DATA SET)
MODEL INPUT: INTEREST SCORE
MODEL OUTPUT: PROBABILITY FOR CORRESPONDING ATTRIBUTE
• VALIDATE MODEL
• CONTINUOUSLY EVALUATE AND PERIODICALLY UPDATE MODELS

FIG. 2

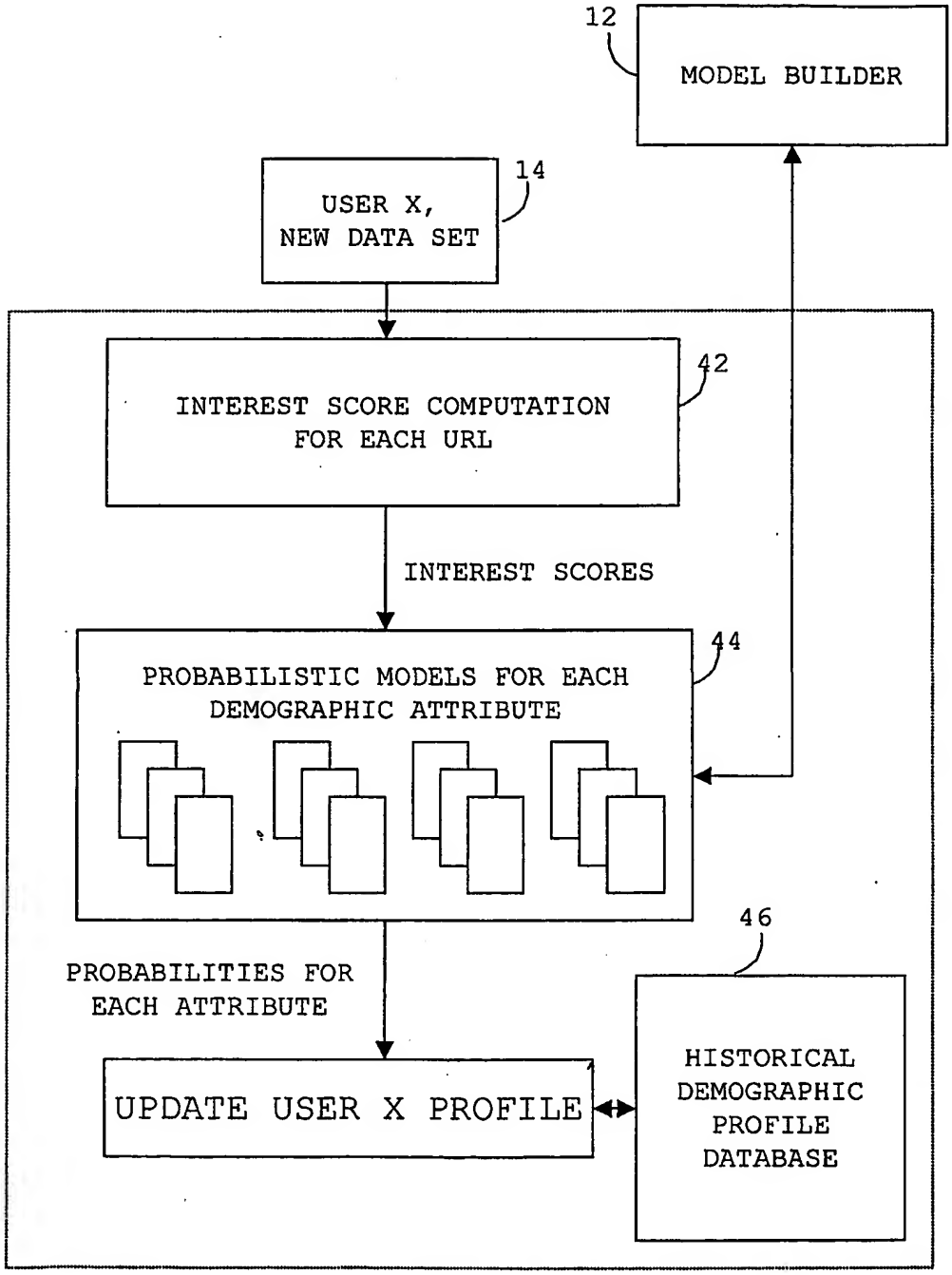
12
MODEL BUILDER
USER X,
NEW DATA SET
14
INTEREST SCORE COMPUTATION
FOR EACH URL
42
INTEREST SCORES
PROBABILISTIC MODELS FOR EACH
DEMOGRAPHIC ATTRIBUTE
44
46
PROBABILITIES FOR
EACH ATTRIBUTE
UPDATE USER X PROFILE
HISTORICAL
DEMOGRAPHIC
PROFILE
DATABASE

FIG. 3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US01/32178

A. CLASSIFICATION OF SUBJECT MATTER
IPC(7) : G06F 17/60
US CL : 703/6
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
U.S. : 703/6;705/1,7

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Please See Continuation Sheet

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 6,112,186 A (BERGH et al.) 29 August 2000 (29.08.2000).	1-86
Y,P	US 6,236,978 B1 (TUZHILIN) 22 May 2001 (22.05.2001).	1-86
Y	US 6,119,101 A (PECKOVER) 12 September 2000 (12.09.2000).	1-86
Y	US 5,490,060 A (MALEC et al.) 06 February 1996 (06.02.1996).	1-86
Y	US 5,983,222 A (MORIMOTO et al.) 09 November 1999 (09.11.1999).	1-86
Y,P	US 6,163,773 A (KISHI) 19 December 2000 (19.12.2000).	1-86
Y	US 5,819,245 A (PETERSON et al.) 06 October 1998 (06.10.1998).	1-86
Y	US 6,128,608 A (BARNHILL) 03 October 2000 (03.10.2000).	1-86
Y,P	US 6,157,921 A (BARNHILL) 05 December 2000 (05.12.2000).	1-86

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search
26 January 2002 (26.01.2002)

Date of mailing of the international search report
15 FEB 2002

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231
Facsimile No. (703)305-3230

Authorized officer
James P. Trammell Peggy Harrod
Telephone No. (703) 305-3900

Form PCT/ISA/210 (second sheet) (July 1998)

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US01/32178

Continuation of B. FIELDS SEARCHED Item 3:
USPAT, US-PGPUB, EPO, JPO, DERWENT, IBM TDB, Dialog, ACM, IEEE search terms: demographic profile, Bayes, internet, neural network.